Determination of most stable housekeeping genes, differentially regulated target genes and sample integrity: *BestKeeper© - Excel spreadsheet tool using a Repeated Pair-wise Correlation and Regression Analysis*

Michael W. Pfaffl

Institute of Physiology, Center of Life and Food Sciences, TUM - Weihenstephan, Germany *pfaffl@wzw.tum.de*

INTRODUCTION

A biologically meaningful reporting of target mRNA copy numbers requires accurate and relevant normalisation. The reliability of any real-time RT-PCR experiment can be improved by including invariant endogenous controls in the assay to correct for sample to sample variations in RT-PCR efficiency and errors in sample quantification. Specific errors in the quantification of mRNA transcripts are easily compounded with any variation in the amount of starting material between the samples, e.g. caused by sample-to-sample variations, variations in RNA integrity, RT efficiency differences and cDNA sample loading variations. ***But the quality of normalized quantitative expression data cannot be better than the quality of the normalizer itself! Any variation in the normalizer will obscure real changes and produce artifactual results.***

The basis of data normalisation is the expression result of an endogenous desirable unregulated reference gene transcript. Here a central questions arise: ***What is the appropriate reference gene in my experimental treatment and investigated tissue?*** Commonly used housekeeping genes, e.g. GAPDH, albumin, actins, tubulins, cyclophilin, micro-globulins, 18S rRNA or 28S rRNA are suitable, since they are present in all nucleated cell types, necessary for basic cell survival and considered to be stable in various tissues. But numerous treatments and studies have already shown that the mentioned references are regulated and vary under experimental conditions. If a desired reference gene is regulated in a specific experimental trial it remains to the investigator to decide which gene can fit the hypothesis of a non regulated reference for a reliable normalisation procedure. Therefore more than one housekeeping genes must be tested and the expression results must be combined to an weighted expression index.

Development of *BestKeeper©*

Therefore, a new Excel based software tool was established, named *BestKeeper©*. It determines the best suited housekeeping genes, out of ten candidates. All data processing is based on crossing point (CP) or alternatively on threshold cycle (Ct) basis. The usage of the CP or Ct seems to be the best estimator for the expression levels, because is normal distributed and parametric test can be performed in correlation and regression analysis.

The crossing point datasets are compared over the entire experimental study. No differentiation between treated groups are made. Out of the *'best suited housekeepers'* the software computes their geometric mean called *BestKeeper Index,* according to earlier optimisation calculations made by Vandesompele et al. (*Gen Biol,* 2002).

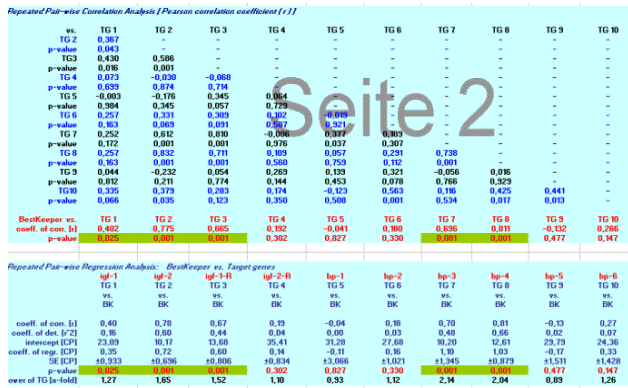
$$\text{BestKeeper Index} = \sqrt[n]{CP_1 \times CP_2 \times CP_3 \times \ldots\ldots \times CP_n}$$

Further data processing and the exclusion of candidate housekeepers is based on a ***Repeated Pair-wise Correlation Analysis and Regression Analysis***. The expression level variation in different samples helps the researcher to decide which gene to include or exclude from the *BestKeeper Index* calculation.

Further on, up to ten target genes are studied and underwent a pair wise *Repeated Pair-wise Correlation Analysis* and *Regression Analysis* versus the calculated *BestKeeper Index,* similar to procedure of the housekeepers. It allow an exact classification of the target gene expression pattern, either close related to the selected housekeepers or differentially expressed. If expressed stable, like the *BestKeeper Index*, they can be considered to be integrated into the index estimation. If the regression analysis of a distinct target genes results in low significance, then it is differentially expressed from the housekeepers.

Repeated Pair-wise Regression Analysis: BestKeeper vs. Housekeping genes

	Ubi HKG 1 vs. BK	GAPDH HKG 2 vs. BK	beta-Actin HKG 3 vs. BK	18 S HKG 4 vs. BK	HKG 5 vs. BK
coeff. of corr. [r]	0,77	0,82	0,83	0,90	
coeff. of det. [r^2]	0,59	0,68	0,69	0,81	
intercept [CP]	8,71	8,49	5,16	-13,70	
coeff. of regr. [CP]	0,67	0,72	0,73	1,48	
SE [CP]	±0,67	±0,591	±0,576	±0,839	
p-value	0,001	0,001	0,001	0,001	
power of HKG [x-fold]	1,60	1,65	1,66	2,79	



Verification of BestKeeper

To verify the *BestKeeper©* software tool the expression of 14 genes were investigated in 31 bovine Corpora lutea, partly treated with Prostaglandin (PG) F2α analogue Cloprostenol (Estrumate®). **4 housekeeping genes** with an assumed stable expression were investigated: **Ubiquitin (UBQ), glyceraldehyd-3-phosphate dehydrogenase (GAPDH), β-actin** and **18S ribosomal subunit**.

Further the expression of **10 target genes, all members of the somatotropic axis,** were assembled: **IGF-1** (insulin-like growth factors type 1), **IGF-2, IGFR-1** (insulin-like growth factor receptor type 1), **IGFR-2, IGFBP-1 to 6** (IGF binding protein type 1 to 6).

Analysis of sample integrity and expression stability within HKGs

Since the occurrence of outliers among prepared samples can obscure the accuracy of the estimation, individual sample preparations are tested for their integrity (e.g. cDNA quantity and quality) as well as their expression stability. An **intrinsic variance (InVar)** of expression for a single sample is calculated as a meaned squared difference of single sample's CP value for one factor from a mean CP value of the same factor.

$$\text{InVar}_m\,[\pm\text{CP}] = \frac{1}{n-1}\sum_{i=1}^{n}\left(CP_n^m - meanCP_n\right)^2$$

The researcher can decide which individual sample can be **removed from the index calculation**, because of high and significant variations in expression levels compared to others.

For a following data processing and determination of treatment differences in the experimental groups the CP or Ct datasets can be imported into further analysis software such as ***REST*** (Relative Expression Software Tool, Pfaffl et al., *NAR* 2002) or ***Q-Gene*** (Muller et al., *BioTechniques* 2002).

Conclusion

In the presented *BestKeeper©* software application up to ten housekeepers and ten target genes can be studied by usage of a pair wise *Repeated Pair-wise Correlation Analysis* and *Regression Analysis*. Under a specific experimental treatment an easy determination of similar or outstanding HKG or TG expression profiles can be made. According to the ***BestKeeper Index* that is based on the expression data of at least three housekeeping genes, a more robust basis for normalisation of real-time PCR can be postulated.** Further the target genes are compared with the index itself and allow an exact classification of the expression pattern, either comparable to the investigated housekeepers or differentially expressed.

Expression stability of the individual sample template integrity (e.g. cDNA quantity, cDNA quality and expression stability) can be determined by an intrinsic variation analysis.

The latest version of *BestKeeper©* can be downloaded free of charge from **http://www.wzw.tum.de/gene-quantification/bestkeeper.html**